Noname manuscript No.
(will be inserted by the editor)

Geometrical correlation indices using homological constructions on manifolds

Alberto J. Hernández · Maikol Solís · Ronald A. Zúñiga-Rojas

the date of receipt and acceptance should be inserted later

Abstract The course of dimensionality is a common problem in statistics and data analysis. Variable sensitivity analysis methods are a well studied and established set of tools designed to overcome these sorts of problems. However, as this work shows, these methods fail to capture relevant features and patterns hidden within the geometry of the enveloping manifold projected into a variable. We propose an index that captures, reflects and correlates the relevance of distinct variables within a model by focusing on the geometry of their projections. The analysis was made with an original R-package called TopSA, short for Topological Sensitivity Analysis. The TopSA R-package is available on the site https://github.com/maikol-solis/TopSA.

Keywords Betti numbers · Data analysis · Homology · Topological manifolds · Sensitivity Analysis · Simplexes

Mathematics Subject Classification (2010) 49Q12 · 52C35 · 57T30 · 55N35

1 Introduction

Let $(X_1, X_2, \dots X_p) \in \mathbb{R}^p$ for $p \geqslant 1$ and $Y \in \mathbb{R}$ two random variables. Define the non-linear regression model as

$$Y = m(X_1, X_2, \dots X_p) + \varepsilon. \tag{1}$$

Here ε is a random noise independent of $(X_1, X_2, \dots X_p)$. The unknown function $m : \mathbb{R}^p \mapsto \mathbb{R}$ describes the conditional expectation of Y given $(X_1, X_2, \dots X_p)$.

Alberto J. Hernández
Centro de Matemática Pura y Aplicada (CIMPA), Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica.

Maikol Solís
Centro de Matemática Pura y Aplicada (CIMPA), Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica. **Corresponding author:** E-mail: maikol.solis@ucr.ac.cr

Ronald A. Zúñiga-Rojas
Centro de Investigaciones Matemáticas y Meta-Matemáticas (CIMM), Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica.

Suppose as well that $(X_{i1}, X_{i2} \dots X_{ip}, Y_i)$ for $i = 1, \dots, n$ is a size n sample for the random vector$(X_1, X_2 \dots X_p, Y)$.

If $p \gg n$, the model (1) suffers from the *"curse of dimensionality"*, a term introduced by Bellman (1957) and Bellman (1961), where he showed that the sample size n required to fit a model increases with the number of variables p. In a statistical context, the model selection techniques solve the problem using indicators as the AIC or BIC or more advanced techniques as Ridge or Lasso regression. The interested reader can find a comprehensive survey of these methodologies in Hastie et al. (2009).

Professionals on computational modelling deal with these problems all the time; the choosing of relevant variables is a recurring task for them. Another way to approach this problem is through the variable importance measures or indices. These indices rely on indicators to find the relevance or sensitivity of the variables in a model. The works of Saltelli et al. (2002), Saltelli et al. (2007), Saltelli et al. (2009) and Wei et al. (2015) compile different approaches to estimate those indicators. They present techniques based on differences methods; parametric and non-parametric settings, variance-based measures, moment independent measures, and graphical techniques.

These techniques overlook the geometric arrangement of the data to build the complete relation between X and Y and then present that information in the form of an indicator. Depending on this simplification they do not consider the geometric properties of the data. For example, most indices will fail to recognize structure when the input variable is of zero-sum, treating it as random noise.

The analysis of topological data is a recent field of research that aims to overcome these shortcomings. Given a set of points generated in space, it tries to reconstruct the model through an embedded manifold that covers the data set. With this manifold, we can study the characteristics of the model using topological and geometrical tools instead of using classic statistical tools.

Two classic tools used to discover the intrinsic geometry of the data are the Principal Components Analysis (PCA) and the Multidimensional Scaling (MDS). The PCA transforms the data in a smaller linear space preserving the statistical variance. The other approach, the MDS, performs the same task but preserving the distances between points. Recent methods like the isomap algorithm developed by Tenenbaum (2000) and expanded by Bernstein et al. (2000); Balasubramanian (2002) unify these two concepts to allow the reconstruction of a low-dimensional variety for non-linear functions. Using the geodesic distance it identifies the corresponding manifolds and search lower dimension spaces to project it.

In recent years, new theoretical developments use tools such as persistent homology, simplicial complexes, and Betti numbers to reconstruct manifolds, the reconstruction works for clouds of random data and functional data, see Ghrist (2008), Carlsson (2009, 2014). In Gallón et al. (2013) and in Dimeglio et al. (2014) some examples are presented. This approach allows handling *" Big Data"* quickly and efficiently, for example, see Snášel et al. (2017).

In this work, we aim to connect the concepts of sensitivity analysis with the analysis of topological data through a geometrical correlation index. By doing this, it will be possible to create relevance indicators for statistical models using the geometric information extracted from the data.

The outline of this paper is: Section 2 deals with basic notions, both in sensitivity analysis and in topology. In Subsection 2.1 some of the most used and well-

known statistics methods are reviewed and commented. We finish this subsection with an example that motivates this work. Subsection 2.2 deals with preliminaries in homology. Section 3 explains the method used to create our sensitivity index; Subsection 3.1 describes the construction of the neighborhood graph, and deals with different topics such as *the importance of scale*, the *Ishigami Model* and presents a strip of code to determine the radius of proximity. Subsection 3.2 describes the algorithm used to construct the homology complex through the *Vietoris-Rips complex* and Subsection 3.3 explains our proposed sensitivity index. Section 4 contains a description of our results, it describes the software and packages used to run our theoretical examples. Subsection 4.1 is a full description of each theoretical example together with a visual aid, such as graphics and tables describing the results. Subsection (4.2) is an application of our algorithm to a well known hydrology model. Finally, Section 5 contains our conclusions and explores scenarios for future research.

2 Preliminary Aspects

2.1 Sensitivity Analysis

According to Saltelli et al. (2009), the sensitivity analysis (SA) is *"the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input"*. The process of validation could be seen as an optional step. But, a complete analysis requires it to find hidden patterns and uncertainties in the data and to identify relevant factors and validating simplifications of the problem.

We could classify sensitivity analysis methods into one-at-the-time and global sensitivity analysis. In the one-at-the-time methods, one variable is sampled at a time while keeping the rest constant. The methods reveal the pattern of the model, however; they work if the problem is linear. Otherwise, features of the data can be missed and misleading conclusions can arise. The correlations and regression sensitivity analysis share this problematic behavior. We refer the reader to Saltelli et al. (2007) for a further review on the flaws and shortcomings of linear sensitivity analysis models.

The standarized regression coefficients are a basic tool to detect linear sensitivities between the inputs and the outputs. It relies in the multiple linear regression,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

Set $\mathbf{X} = [\mathbf{1}, X_1, \ldots, X_p]$ where $\mathbf{1}$ is a vector of ones and the variables X_i's are disposed as columns. We know that the least-square estimator for the β's are

$$\begin{aligned} (\hat{\beta}_0, \ldots, \hat{\beta}_p)^\top &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y \\ Y &= \mathbf{X}(\hat{\beta}_0, \ldots, \hat{\beta}_p)^\top. \end{aligned} \tag{2}$$

The regression coefficients $\hat{\beta}_i$ $(i = 1, \ldots, p)$ measure how much sensitive are the X_i's with respect Y. However, they fail in describe the relative importance between

the variables because their values are still in the units of the inputs variables. To solve this, we can rewrite the fitted regression model 2 as

$$\frac{\hat{Y}-\bar{Y}}{\hat{s}} = \sum_{i=1}^{n} \left(\frac{\hat{\beta}_i \hat{s}_i}{\hat{s}} \right) \frac{(X_i - \bar{X}_i)}{\hat{s}_i},$$

where

$$\hat{s} = \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})/(n-1)}, \quad \hat{s}_i = \sqrt{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)/(n-1)}$$

are the standard deviations for the output samples and the input samples respectively. The values $\rho_i^{\mathrm{SRC}} = \hat{\beta}_i \hat{s}_i / \hat{s}$ are called the standard regression coefficients (SRC) of X_i with respect Y. It reflects the sensitiviy of X_i removing any units in the data. So the larger $|\rho_i^{\mathrm{SRC}}|$ the most influential is X_i to the model.

To discover the nonlinear patterns of the data one could use other techniques. Some popular and well-documented ones are screening, Morris method, global variance or the moment independent measures. For instance, the Morris method creates an approximation of the variation over the variable we want to study. This method can detect no linearity, but due to its nature assumes monotonicity. This characteristic is not always present in complex models Hamby (1994). The higher-order effects are difficult to estimate due to their demanding computational requirements Saltelli et al. (2007).

One popular method to asses relevant variables is the variance-based global sensitivity analysis. The method was proposed by Sobol' (1993) based on the ANOVA decomposition. He proved that if f is an squared integrable function, then he decomposed it in the unitary cube as:

$$Y = f = f_0 + \sum_i f_i + \sum_{\substack{ij \\ i \neq j}} f_{ij} + \cdots + f_{12\cdots p} \tag{3}$$

where each term is also square integrable over the domain. Also, each function is defined by $f_i = f_i(X_i)$, $f_{ij} = f_{ij}(X_i, X_j)$ and so on. This decomposition has 2^p terms and the first one, f_0, is constant. The remaining are non-constant functions. Sobol also proved that this representation is unique if each term has zero mean and the functions are pairwise orthogonal. Equation (3) could have been seen as the decomposition of the output variable Y into its effects due to the interaction of none, one or multiple variables. Taking expectation in Equation (3) and simplifying the expression we get:

$$\begin{aligned} f_0 &= \mathbb{E}[Y] \\ f_i(X_i) &= \mathbb{E}[Y|X_i] - \mathbb{E}[Y] \\ f_{ij}(X_i, X_j) &= \mathbb{E}[Y|X_i, X_j] - f_i - f_j - f_0 \end{aligned}$$

and so on for all combinations of variables.

Once with this orthogonal decomposition, we measure the variance of each element. The global-variance method estimates the regression curves (surfaces) for each dimension (or multiple dimensions) removing the effects due to variables in

lower dimensions. Then, it gauges the variance of each curve (surface) normalized by the total variance in the model. For the first and second-order effects the formulas remain as:

$$S_i = \frac{\mathrm{Cov}(f_i(X_i), Y)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)} \tag{4}$$

$$S_{ij} = \frac{\mathrm{Cov}(f_{ij}(X_i, X_j), Y)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i, X_j])}{\mathrm{Var}(Y)} - S_i - S_j. \tag{5}$$

and so on for the higher order terms.

To quantify how much we could apportion of each variable to the whole model, the total effect is estimated by

$$S_{T_i} = 1 - \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\sim i}])}{\mathrm{Var}(Y)}$$

The moment independent indices estimates relevant inputs in other way. These estimate the average of the distance between the random variable Y and $Y|X_i = x$ for any $x \in \mathbb{R}$. The initial ideas come from Borgonovo et al. (2014) which used monotonic invariant transformations to gauge the distance between two probability measures. Therefore, if the variable X_i is irrelevant or independent to the model, the random variables Y and $Y|X_i = x$ will be almost identical and the impact will be small. Otherwise the impact will be significant. The following are popular measures using this technique:

- **Kolmogorov-Smirnov:** $\mathbb{E}[\sup_{y\in\mathbb{R}} |F_Y(y) - F_{Y|X_i}(y)|]$,
- **Radon-Nykodym:** $\frac{1}{2}\int |f_Y(y) - f_{Y|X_i}(y)|dy$,
- **Kullback-Leiber:** $\frac{1}{2}\int f_Y(y) \log\left(\frac{f_Y(y)}{f_{Y|X_i}(y)}\right) dy$

where F_Y and $F_{Y|X_i}$ represent the distribution function for Y and $Y|X_i$ respectively. The density functions are presented as f_Y and $f_{Y|X_i}$.

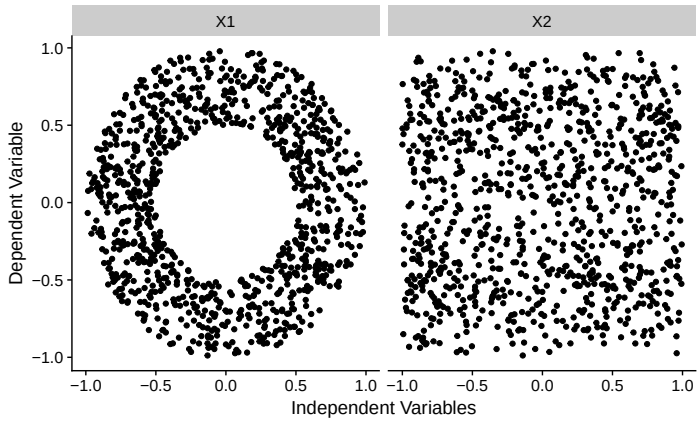
All those methods try to compact all the features of the space in a single measure and then average over the set of values into the sample.

However, the sample could have particularities not detected by the method. For example, Figure 1 bellow presents a toy example where the data points are arranged into a circle with a hole in the middle (further details in Section 4). The first variable presents all the information about the model while the second one is a noisy arrangement.

We estimated the SRC for both cases. Here, both results agree that the variables X_1 and X_2 are irrelevant to the model. In this paper we propose an algorithm that handles this examples and notices such geometric and topological details within the data.

2.2 Homology

In this subsection we recall some of the topological definitions needed to understand the frame of our work.



	X_1	X_2
ρ^{SRC}	0.00	-0.01

Fig. 1: Estimation of correlation index for the example of a Circle with a hole (see Section 4.1. In this case both variables are irrelevant to the model.

n*—*topological manifold is a Hausdorff space M with a countable basis such that each point $x \in M$ has an open neighborhood that is homeomorphic with an open subset of $\mathbb{R}^n$.

We recall here that *Hausdorff* means that any two distinct points lie in disjoint open subsets. A *countable basis* means countable family of open subsets such that each open subset is the union of a subfamily. The reader may see Barden and Thomas (2003) for details.

In terms of differential manifolds, a *chart* on a topological manifold M is a pair (U, ϕ) where $U \subseteq M$ is an open subset and $\phi : M \to \mathbb{R}^n$ is a homeomorphism that sends U to an open subset $V = \phi(U) \subseteq \mathbb{R}^n$. The set $U \subseteq M$ is so-called a *coordinate neighborhood* and the set $V \subseteq \mathbb{R}^n$ is so-called its *coordinate space*. An *atlas* for a topological manifold M is a collection of charts $\mathcal{A} = \{(U_i, \phi_i) | i \in I\}$ such that the coordinate neighborhoods of $\mathcal{A}$ cover the whole manifold M:

$$\bigcup_{i \in I} U_i = M.$$

To avoid ambiguity, we introduce the *coordinate transformations*: if two distinct coordinate neighborhoods are not disjoint then, we can define the following functions from the images of the overlaping open subset:

$$\theta_{12} = \phi_1 \circ \phi_2^{-1} : \phi_2(U_1 \cap U_2) \to \phi_1(U_1 \cap U_2)$$

and

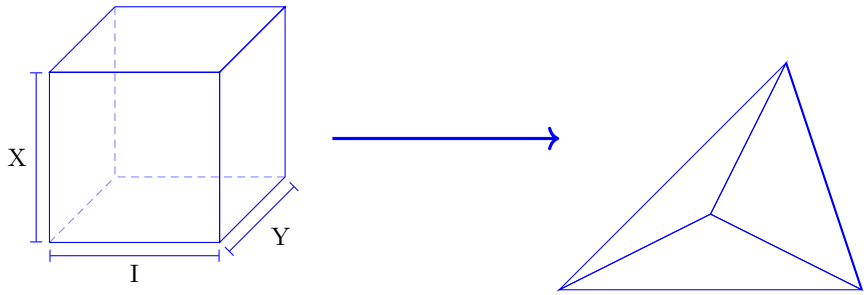
$$\theta_{21} = \phi_2 \circ \phi_1^{-1} : \phi_1(U_1 \cap U_2) \to \phi_2(U_1 \cap U_2)$$

From Barden & Thomas Barden and Thomas (2003) An ***n*—*differential manifold*** is a topological manifold M of dimension n together with a maximal smooth atlas on it. The reader may consult Lang (1963) for differential manifolds details.

To define a simplex, let consider the unit interval $I = [0, 1]$ and two topological spaces X and Y. Consider the triple product $X \times Y \times I$ and the quotient $(X \times Y \times I)/\sim$ where the relation $\sim$ is given by the identifications

$$(x, y_0, 0) \sim (x, y_1, 0) \quad \text{and} \quad (x_0, y, 1) \sim (x_1, y, 1).$$

In other words, we are collapsing the subspaces $X \times Y \times \{0\}$ and $X \times Y \times \{1\}$ to X and Y respectively. To visualize it, take for instance $X = Y = I$ two real closed intervals. Hence, we are collapsing two opposite faces of a cube onto line segments so that, the cube becomes a tetrahedron:



According to Hatcher (2000) (may see also Munkres (1984)), a $\triangle$-complex is a generalization of a simplex. Recall that 0-simplices are points, called *vertices*. As well as 0-simplices, we have line segments as 1-simplices, called *edges*, 2-simplices called *faces*, 3-simplices called *tetrahedron*, and a generalization of dimension n will be a convex set in $\mathbb{R}^m$ containing $\{v_0, v_1, \ldots, v_n\}$ a subset of $n + 1$ distinct points that do not lie in the same hyper plane of dimension n or, equivalently, that the vectors $\{w_j = v_j - v_0\}$ are linearly independent. In such a case, we are denoting the points $\{v_0, v_1, \ldots, v_n\}$ as vertices, and the usual notation would be $[v_0, v_1]$ for edges, $[v_0, v_1, v_2]$ for faces, $[v_0, \ldots, v_4]$ for tetrahedron, and $[v_0, \ldots, v_n]$ for *n-simplicies*.

A $\triangle$-complex is a qoutient topological space of a collection of disjoint simplices identifying some of their faces via certain homeomorphisms $\{\alpha\}_{\alpha \in \mathcal{A}}$ that preserve the order of the vertices. Those homeomorphisms are linear, and give us a better notation for n-simplices: e^n_α.

Now, we may define the simplicial homology groups of a $\triangle$-complex X as follows. Let consider the free abelian group $\triangle_n(X)$ with open n-simplices $e^n_\alpha \subseteq X$ as basis elements. It means that the elements of this group $\triangle_n(X)$, known as *chains*, look like linear combinations of the form

$$c = \sum_\alpha n_\alpha e^n_\alpha \tag{6}$$

with integer coeficients $n_\alpha \in \mathbb{Z}$. We also could wirte chains as linear combinations of characteristic maps

$$c = \sum_\alpha n_\alpha \sigma_\alpha \tag{7}$$

where every $\sigma_\alpha \colon \triangle^n \to X$ is the corresponding characteristic map of each e^n_α, with image the closure of e^n_α. So, $c \in \triangle_n(X)$ is a finite collection of n-simplices in X with integer multiplicities n_α. The boundary of the n-simplex $[v_0, \ldots, v_n]$ consists

of the various $(n-1)$-simplices $[v_0,\ldots,\hat{v}_j,\ldots,v_n]=[v_0,\ldots,v_{j-1},v_{j+1},\ldots,v_n]$. For chains, the boundary of $c=[v_0,\ldots,v_n]$ is an oriented $(n-1)$-chain of the form

$$\partial c=\sum_{j=0}^{n}(-1)^j[v_0,\ldots,\hat{v}_j,\ldots,v_n] \tag{8}$$

which is a linear combination of faces. This allows us to define the *boundary homomorphisms* for a general $\triangle$-complex X, $\partial_n\colon\ \triangle_n(X)\to\triangle_{n-1}(X)$ as follows:

$$\partial_n(\sigma_\alpha)=\sum_{j=0}^{n}(-1)^j\sigma_\alpha|[v_0,\ldots,\hat{v}_j,\ldots,v_n]. \tag{9}$$

Hence, we get a sequence of homomorphisms of abelian groups

$$\cdots\to\triangle_n(X)\xrightarrow{\partial_n}\triangle_{n-1}(X)\xrightarrow{\partial_{n-1}}\triangle_{n-2}(X)\to\cdots\to\triangle_1(X)\xrightarrow{\partial_1}\triangle_0(X)\xrightarrow{\partial_0}0$$

where $\partial\circ\partial=\partial_n\circ\partial_{n-1}=0$ for all n. This is usually known as a *chain complex*. Since $\partial_n\circ\partial_{n-1}=0$, the $\mathrm{Im}(\partial_n)\subseteq\ker(\partial_{n-1})$, and so, we define the n^{th} *simplicial homology group* of X as the quotient

$$h_n^{\triangle}(X)=\frac{\ker(\partial_n)}{\mathrm{Im}(\partial_{n+1})}. \tag{10}$$

The elements of the kernel are known as *cycles* and the elements of the image are known as *boundaries*.

Easy computations of sequences give us the simplicial homology of some examples: the circle $X=\mathbb{S}^1$:

$$H_0^{\triangle}(\mathbb{S}^1)\cong\mathbb{Z},\quad H_1^{\triangle}(\mathbb{S}^1)\cong\mathbb{Z},\quad H_n^{\triangle}(\mathbb{S}^1)\cong 0\quad\text{for}\quad n\geqslant 2,$$

and the torus $X=T\cong\mathbb{S}^1\times\mathbb{S}^1$:

$$H_0^{\triangle}(T)\cong\mathbb{Z},\quad H_1^{\triangle}(T)\cong\mathbb{Z}\oplus\mathbb{Z},\quad H_2^{\triangle}(T)\cong\mathbb{Z},\quad H_n^{\triangle}(T)\cong 0\quad\text{for}\quad n\geqslant 3.$$

In a very natural way, one can extend this process to define singular homology groups $H_n(X)$. This process, nevertheless, is not trivial but natural. If X is a $\triangle$-complex with finitely many n-simplices, then $H_n(X)$ (and of course $H_n^{\triangle}(X)$) is finitely generated. The n^{th}*-Betti number* of X is the number of summands isomorphic to the additive group $\mathbb{Z}$. The reader may see Hatcher (2000) or Munkres (1984) for details.

Definition 1 (VR neigboorhood graph) Given $S\subseteq\mathbb{R}^p$ and scale $\varepsilon\in\mathbb{R}$, the VR neighborhood graph is a graph where $G_\varepsilon(V)=(V,E_\varepsilon(V))$ and

$$E_\varepsilon(V)=\{\{u,v\}\mid d(u,v)\leqslant\varepsilon,u\neq v\in V\}.$$

Definition 2 (VR expansion) Given a neigborhood graph G, their Vietoris-Rips complex $\mathcal{V}(G)$ is defined as all the edges of a simplex σ that are in G. In this case σ belongs to $\mathcal{V}(G)$. For $G=(V,E)$, we have

$$\mathcal{V}(G)=V\cup E\cup\left\{\sigma\mid\binom{\sigma}{2}\subseteq E\right\}.$$

where σ is a simplex of G.

3 Methodology

Recall the model (1). Here the random variables $(X_1, \ldots, X_p)$ are distorted by the function m and its topology. Our aim is to measure how much each of the X_i influenced this distortion, i.e, we want to determine which variables influence the model the most.

In Section 2.1 we reviewed the most common statistical indices to estimate the sensitivity or correlation of X_i for $i = 1, \ldots, p$ with respect to Y. In our case, we want to consider the geometry of the point-cloud, its enveloping manifold and create an index that will reveal information about the model.

The first step is to create a neighborhood graph for the point-cloud formed by (X_i, Y) where an edge is set if a pair of nodes are within an ε of euclidean distance. In this way, we connect only the nodes nearby by a fixed distance. With this neighborhood graph, we construct the persistent homology using the method of Zomorodian (2010) for the Vietoris-Rips (VR) complex.

The algorithm of Zomorodian (2010) works in two-phases: First it creates a VR neighborhood graph (Definition 1) and then builds, step by step, the VR complex (Definition 2).

These definitions state the procedure to construct the two-phase scheme for the Vietoris-Rips complex with resolution ε:

1. Using Definition 1 compute the neighborhood graph $G_\varepsilon(V)$ with parameter ε.
2. Using Definition 2 compute $\mathcal{V}(G_\varepsilon(V))$. From now on set $\mathcal{V}_\varepsilon(V) = \mathcal{V}(G_\varepsilon(V))$.

This scheme provides us with a geometrical skeleton for the data cloud points with resolution ε. If ε is large there will be more edges connecting points. Here, we could have a large interconnected graph with little information. Otherwise, if ε is small, there would be fewer edges connecting points, resulting in a sparse graph and missing relevant topological features within the data cloud.

In the second step, we unveil the topological structure of the neighborhood graph through the Vietoris-Rips complex. The expansion builds the cofaces related to our simplicial complex. In Section 3.2 we will discuss more about the algorithm used for this purpose.

3.1 Neigborhood graph

The neighborhood graph collects the vertices V, and for each vertex $v \in V$ it adds all the edges $v - -u$ within the set $u \in V$, satisfying $d(v, u) \leqslant \varepsilon$. This brute-force operation works in $O(n^2)$ time.After considering a variety of theoretical examples it becomes clear that the scale factor in the data set is relevant. The scale in one variable may differ with the scale of the output by many orders of magnitude. Thus, proximity is relative to the scale of the axis in which the information is presented and the proximity neighborhood could be misjudged.

The Ishigami model, presented in Figure 2 bellow shows how the proximity neigborhood becomes distorted when scale is taken into consideration.

We conclude that the aspect ratio between both variables define how the algorithm construct the neighborhood graph. Therefore, to use circles to build the neighborhood graph, we need to set both variables into the same scale. The Algorithm 1 constructs the VR-neigborhood graph for a cloud of points with arbitrary scales.

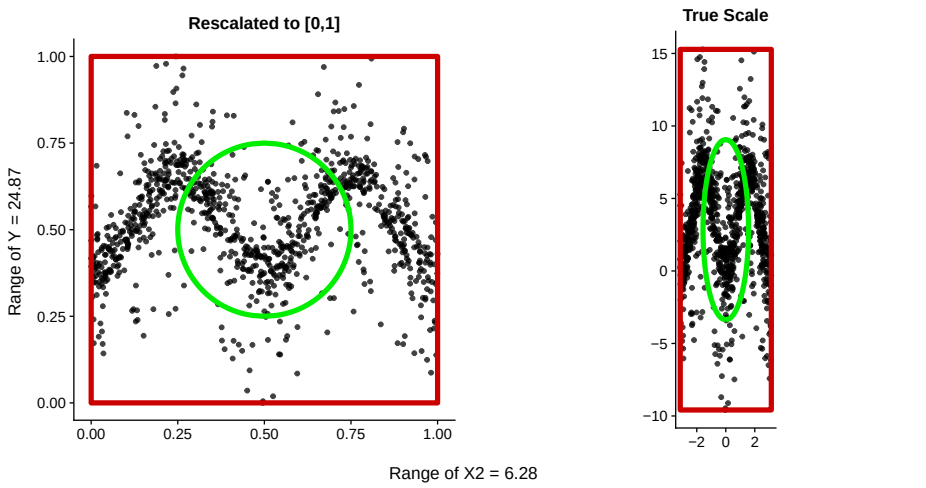


Fig. 2: The Second variable of the Ishigami model escalated to $[0, 1]$ with a circle centered in $(0.5, 0.5)$ and radius 1 (left). The same circle draw in the true scale of the data (right).

Data: A set of points $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$
A value $0 < \alpha < 1$.
Result: The Neigborhood Graph.

Function CREATE-VR-NEIGBORHOOD((X, Y), α)**:**
  $X_r \leftarrow \dfrac{X - \min X}{\max X - \min X}$
  $Y_r \leftarrow \dfrac{Y - \min Y}{\max Y - \min Y}$
  $n \leftarrow \text{length}(Y)$
  DistanceMatrix $\leftarrow$ Matrix$(n \times n)$
  for i = *1:n* **do**
    for j = *1:n* **do**
      DistanceMatrix[i,j] $\leftarrow \sqrt{(X_r[\text{i}] - X_r[\text{j}])^2 + (Y_r[\text{i}] - Y_r[\text{j}])^2}$
    end
  end
  $\varepsilon \leftarrow$ QUANTILE (DistanceMatrix, α)
  AdjacencyMatrix $\leftarrow$ DistanceMatrix $\leqslant \varepsilon$
  NeigborhoodGraph $\leftarrow$ CREATEGRAPH (AdjacencyMatrix, xCoordinates $= X$, y Coordinates $= Y$)
  return (NeigborhoodGraph)
end

Algorithm 1: Procedure to estimate the neigborhood graph from a set of points in the plane.

1. Rescale the points (X_i, Y) $i = 1, \ldots, n$ onto the square $[0, 1] \times [0, 1]$.
2. Estimate the distance matrix between points.
3. With the distance chart, estimate the α quantile of the distances. Declare the radius ε_i as this quantile.
4. Using Definition 1, build the VR-neigborhood graph with ε changed by ε_i for each projection.
5. Rescale the data points to their original scale.

3.2 VR expansion

In his work Zomorodian (2010), Zomorodian describes three methods to build the Vietoris-Rips complex. The first approach builds the complex adding the vertices, the edges and then increasing the dimension to create triangles, tetrahedrons, etc. The second method starts from an empty complex, and adds, step by step, all the simplices stopping in the desired dimension. In the third one, one takes advantage from the fact that the VR-complex is the combinations of cliques in the graph in the desired dimension.

Due to its simplicity, we adopt the third approach and detect the cliques in the graph. We use the algorithm in Eppstein et al. (2010) which is a variant of the classic algorithm from Bron and Kerbosch (1973). This algorithm orders the graph G and then computes the cliques using the Bron-Kerbosch method without pivoting. This procedure reduced the worst-case scenario from time $\mathcal{O}(3^{n/3})$ to $\mathcal{O}(dn3^{d/3})$ where n is the number of vertices and d is the smallest value such that every nonempty subgraph of G contains a vertex of degree at most d.

Constructing the manifold via the VR-complex is not efficient in the sense that the co-faces may overlap, increasing the computational time required., One can overcome this issue by creating an ordered neighborhood graph.

3.3 Geometrical correlation construction

One of the main objectives in a sensitivity analysis is to discover patterns and relationships between the inputs against the output and determine which of those patterns are more influential for the model in consideration. For our case, the patterns are described through empty spaces in the projection space generated by each individual variable. If the point-cloud fills all the domain then the unknown function m applied to X_i produces erratic values of Y. Otherwise, the function sheds an structural pattern which could be recognized geometrically.

The VR-complex $\mathcal{V}(G)$ estimates the geometric structure of the data by filling the voids between close points. Then, we estimate the area of the created object. This number will not give much information about the influence that the variable has within the model. We estimate the area of the minimum rectangle containing all the object. If some input variable presents a weak correlation with the output variable, its behavior will be almost random with uniform distributed points into the rectangular box. In other case, if it has some relevant correlation, it will create a pattern causing empty spaces to appear across the box.

To clarify the notation we will denote as $G_{\varepsilon,i}$ the neighborhood graph generated by the pair of variables (X_i, Y) and the radius ε. Denote as $\mathcal{V}(G_{\varepsilon,i})$ the VR-complex

generated by $G_{\varepsilon,i}$. We also denote as $\text{Area}(\mathcal{V}(G_{\varepsilon,i}))$ as the geometrical area of the object formed by the VR-complex $\mathcal{V}(G_{\varepsilon,i})$.

We define the rectangular box for the projection the data (X_i, Y) as

$$B_i = \Big[\min_{X_i}(\mathcal{V}(G_{\varepsilon,i})), \max_{X_i}(\mathcal{V}(G_{\varepsilon,i}))\Big] \times \Big[\min_{Y}(\mathcal{V}(G_{\varepsilon,i})), \max_{Y}(\mathcal{V}(G_{\varepsilon,i}))\Big].$$

The geometrical area of B_i will be denote as $\text{Area}(B_i)$.

Therefore, we can create the measure,

$$\rho_i^{\text{Geom}} = 1 - \frac{\text{Area}(\mathcal{V}(G_{\varepsilon,i}))}{\text{Area}(B_i)}.$$

Notice that if the areas of the object and the box are similar, then the index ρ_i^{Area} is close to zero. Otherwise, if there is a lot of empty spaces and both areas differ and the index would approach 1.

4 Results

To asses the quality of the estimated index described before, we performed numerical examples. The software used was *R* (R Core Team (2017)), along with the packages `igraph` (Csárdi and Nepusz (2006)) for all the graph manipulations, and the packages `rgeos` and `sp` (Bivand and Rundel (2018); Pebesma and Bivand (2005); Bivand et al. (2013)) for all the geometric estimations. A package containing all these algorithms will be available soon in CRAN.

For all the settings, we sample $n = 1000$ points with the distribution specified in each case. Due to the number of points, we choose the quantile 5% for each case to determine the radius of the neigborhood graph. Further insights about this choosing will be presented in the conclusions section

We will consider five settings, each one with different topological features. The cases are not exhaustive and there are other settings with interesting features as well. However, through this sample we show how the method captures the geometrical correlation of the variables where other classic methods have failed, as well as a case for which our method fails to retrieve the desired information.

4.1 Theoretical examples

The examples considered are the following:

Linear This is a simple setting with

$$Y = 2X_1 + X_2$$

and X_3, X_4 and X_5 independent random variables. We set $X_i \sim \text{Uniform}(-1, 1)$ for $i = 1, \ldots, 5$.

Quartic: This is another simple scheme with

$$Y = X_1 + X_2^4$$

with $X_i \sim \text{Uniform}(-1, 1)$ for $i = 1, 2$.

Circle with hole: The model in this case is

$$\begin{cases} X_1 = r\cos(\theta) \\ Y = r\sin(\theta) \end{cases}$$

with $\theta \sim \text{Uniform}(0, 2\pi)$ and $r \sim \text{Uniform}(0.5, 1)$. This form creates circle with with a hole in the middle.

Connected circles with holes: The model is set in two sections, where in both parts we set $\theta \sim \text{Uniform}(0, 2\pi)$:

1. Circle centered in $(0, 0)$ with radius between 1 and 2:

$$\begin{cases} X_1 = r_1\cos(\theta) \\ Y = r_1\sin(\theta) \end{cases}$$

   where $r_1 \sim \text{Uniform}(1, 2)$.
2. Circle centered in $(1.5, 1.5)$ with radius between 0.5 and 1:

$$\begin{cases} X_1 - 1.5 = r_2\cos(\theta) \\ Y - 1.5 = r_2\sin(\theta) \end{cases}$$

   where $r_2 \sim \text{Uniform}(0.5, 1)$.

Ishigami: The final model is

$$Y = \sin X_1 + 7\ \sin^2 X_2 + 0.1\ X_3^4 \sin X_1$$

where $X_i \sim \text{Uniform}(-\pi, \pi)$ for $i = 1, 2, 3$, $a = 7$ and $b = 0.1$.

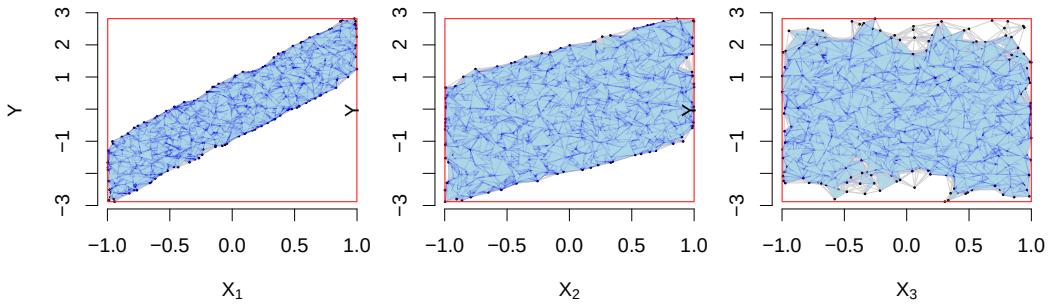
4.1.1 Numerical results

The figures presented in this section represent the estimated manifold for each input variable X_i with respect to the output variable Y. The lower table presents the radius used to build the neighborhood graph, the estimated areas for the manifold object, the reference square and the proposed index.

The linear model in Figure 3 is simple enough to allow us to declare the variable X_1 has the double of relevance compared to variable X_2. The other variables will have less relevant indices. As we expected, the index for X_1 almost doubles the counterpart for X_2. The examples show us how the empty spaces are present according to the relevance level of the variable.

For the Quartic model in Figure 4, we could estimate the theoretical Sobol indices according to Equation (4). The indices are $S_1 = 0.82$, $S_2 = 0.18$ and $S_3 = 0$. We observe in Figure 4 how the Sobol indices match with our algorithm, ranking the variables in order of relevance as X_1, X_2 and X_3.

The model of Circle with a hole in Figure 5 was discussed in the preliminaries. Recall that in this case, both variables were irrelevant to the model, even if the geometric shape showed the contrary. Figure 5 present the results. Observe how the first variable has index equal to 0.48 and the second one with 0.06. These



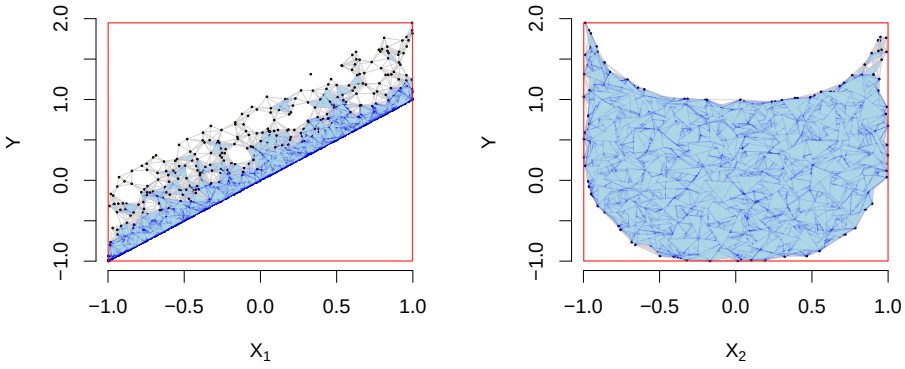
Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	ρ^{Geom}	ρ^{SRC}
X_1	0.08	3.79	11.38	0.67	0.90
X_2	0.11	7.69	11.39	0.32	0.46
X_3	0.12	9.78	11.39	0.14	0.00
X_4	0.12	10.18	11.38	0.10	0.00
X_5	0.12	10.16	11.39	0.11	0.00

Fig. 3: Results for the Linear case.

indices allow us to say the X_1 has an impact into the model much more relevant than X_2.

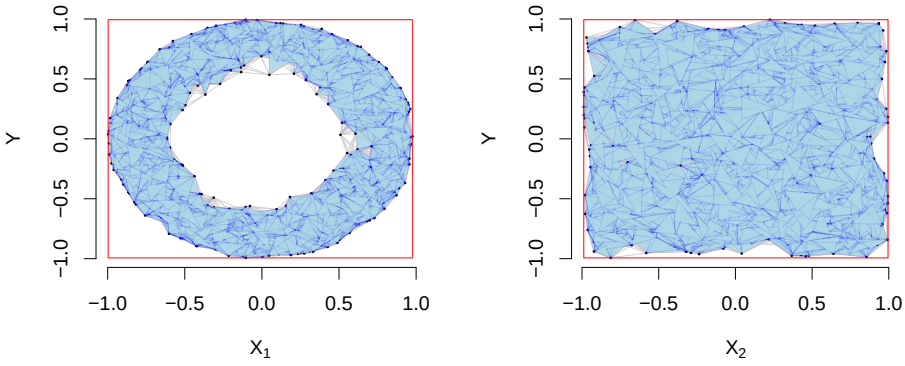
To test our algorithm further, we present the model of Connected circles with holes in Figure 6. Here we created two circles with different scales and positions. Even if the choice is an arguable point for the method, we could capture the most relevant features for each projection. Again, it is possible to rank the variables as X_1 first with index equal to 0.58 and X_2 second with index 0.08.

The final model is produced by the Ishigami function, Figure 7. This is a popular model in sensitivity analysis because presents a strong nonlinearity and non-monotonicity with interactions in X_3. With other sensitivity estimators, the variables X_1 and X_2 have great relevance to the model, while the third one X_3 has almost zero. For a further explanation of this function, we refer the reader to Sobol' and Levitan (1999). In our case, all the variables result into a index near to 0.5. This is relevant because our index has captured the presence of structure, as well as with the model Circle with hole whereas other tools treat them as noise.



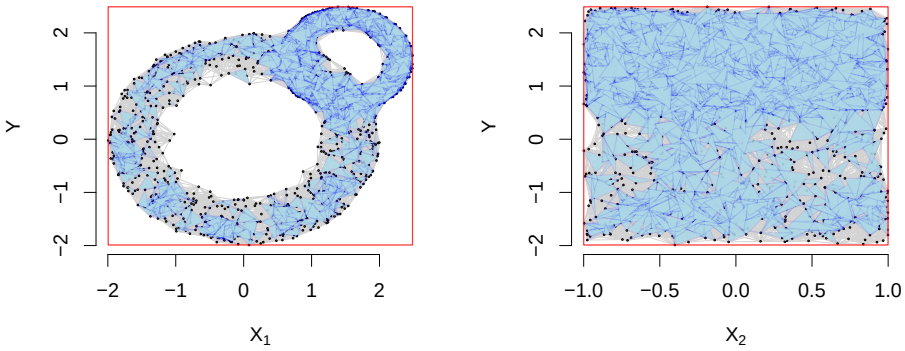
Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	ρ^{Geom}	ρ^{SRC}
X_1	0.06	1.50	5.88	0.75	0.91
X_2	0.11	3.83	5.89	0.35	0.01

Fig. 4: Results for the Quartic case.



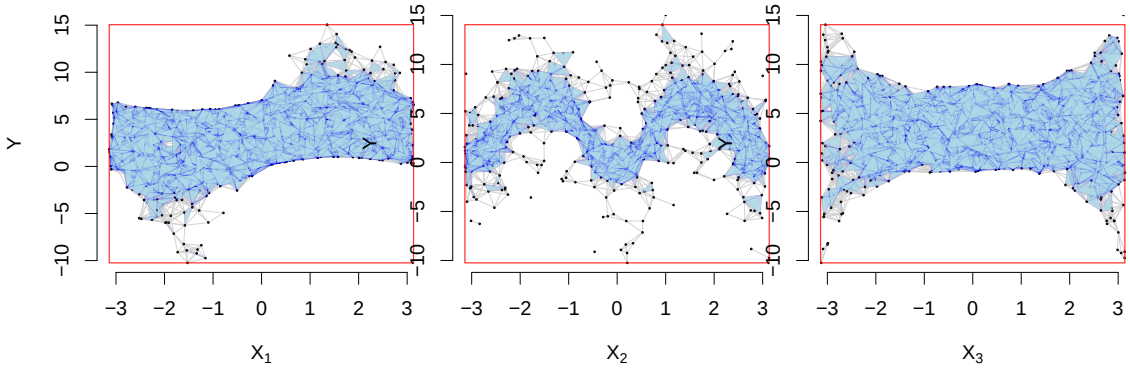
Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	ρ^{Geom}	ρ^{SRC}
X_1	0.11	2.10	3.92	0.46	0.00
X_2	0.13	3.67	3.94	0.07	0.01

Fig. 5: Results for the Circle with 1 hole case.



Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	ρ^{Geom}	ρ^{SRC}
X_1	0.08	9.44	20.07	0.53	0.36
X_2	0.12	8.60	8.96	0.04	0.00

Fig. 6: Results for the Circle with 2 holes case.



Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	ρ^{Geom}	ρ^{SRC}
X_1	0.08	64.83	158.65	0.59	0.44
X_2	0.07	56.78	152.50	0.63	0.02
X_3	0.09	70.96	152.59	0.53	0.01

Fig. 7: Results for the Ishigami case.

4.2 Application: A Hydrology model

One academic real case model to test the performance in sensitivity analysis is the dyke model. This model simplifies the 1D hydro-dynamical equations of Saint Venant under the assumptions of uniform and constant flow rate and large rectangular sections.

The following equations recreate the variable S which measures the maximal annual overflow of the river (in meters) and the variable C_p which is the associated cost (in millions of euros) of the dyke.

$$S = Z_v + H - H_d - C_b \tag{11}$$

with

$$H = \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)$$

$$C_p = \mathbf{1}_{S>0} + \left[0.2 + 0.8 \left(1 - \exp \frac{-1000}{S^4} \right) \right] \mathbf{1}_{S \leqslant 0}$$

$$+ \frac{1}{20} \left(H_d \mathbf{1}_{H_d > 8} + 8 \mathbf{1}_{H_d \leqslant 8} \right) \tag{12}$$

Table 1 shows the inputs ($p = 8$). Here $\mathbf{1}_A(x)$ is equal to 1 for $x \in A$ and 0 otherwise. The variable H_d in Equation (11) is a design parameter for the Dyke's height set as a Uniform$(7, 9)$.

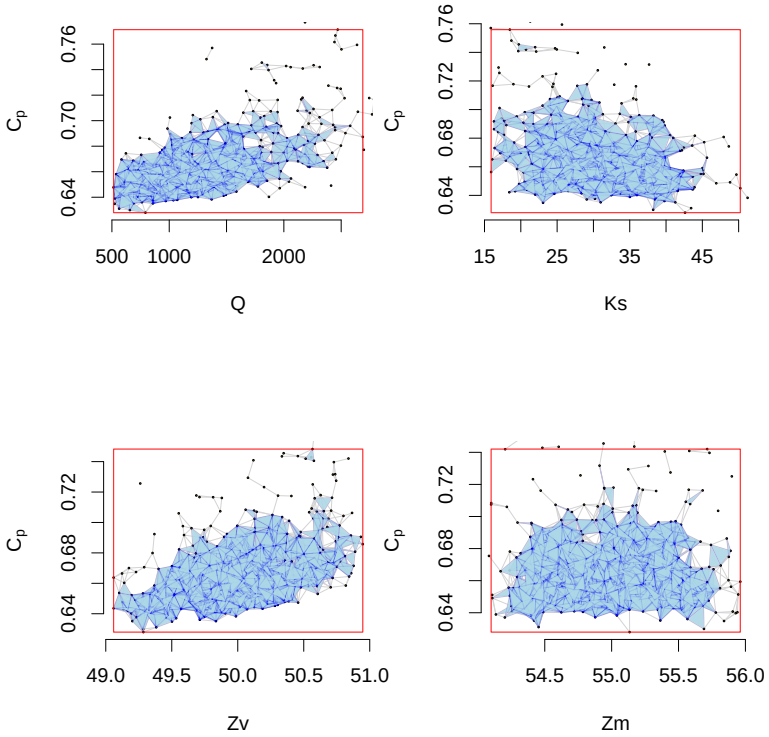
In Equation (4.2), the first term mean a cost of 1 million euros due to a flooding ($S > 0$). The second term corresponds to the cost of the dyke maintenance ($S \leqslant 0$) and the third term is the construction cost related to the dyke. The latter cost is constant for a height of dyke less than 8m and is growing like the dyke height otherwise.

Input	Description	Unit	Probability Distribution
Q	Maximal annual flowrate	m^3/s	Gumbel$(1013, 558)$ truncated on [500, 3000]
K_s	Strickler coefficient	—	$\mathcal{N}(30, 8)$ truncated on $[15, \infty)$
Z_v	River downstream level	m	Triangular$(49, 50, 51)$
Z_m	River upstream level	m	Triangular$(54, 55, 56)$
H_d	Dyke height	m	Uniform$(7, 9)$
C_b	Bank level	m	Triangular$(55, 55.5, 56)$
L	Length of the river stretch	m	Triangular$(4990, 5000, 5010)$
B	River width	m	Triangular$(295, 300, 305)$

Table 1: Input variables and their probability distributions.

For a complete discussion about the model, parameters definition and meaning the reader can review (Iooss and Lemaître, 2015), (de Rocquigny, 2006) and their references.

The work of Iooss and Lemaître (2015) detects the most influential variables for models (4.2) and (11). They use a combination of a Morris screening method, the

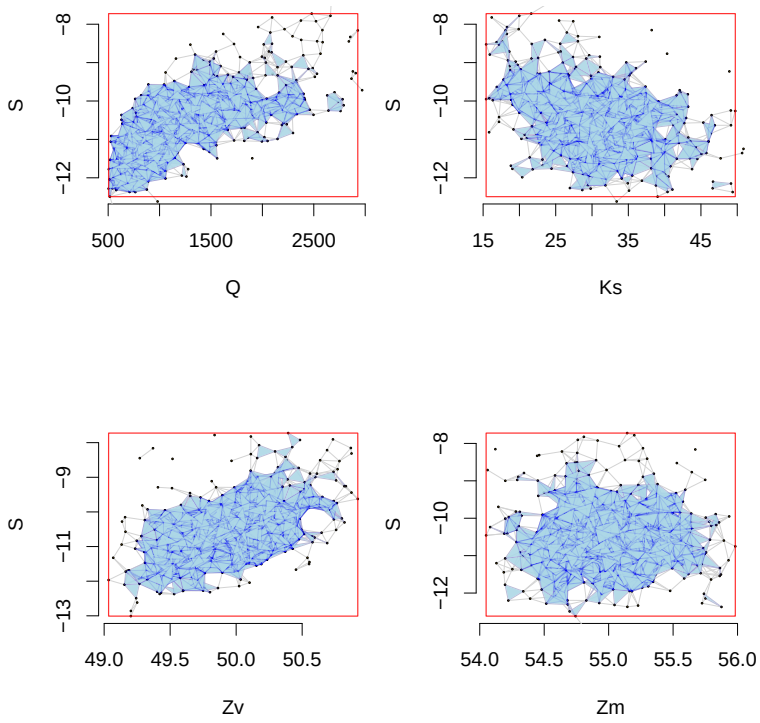


Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	ρ^{Geom}	ρ^{SRC}
Q	0.07	96.81	311.84	0.69	0.63
K_s	0.07	1.58	4.39	0.64	0.38
Z_v	0.08	0.09	0.23	0.61	0.42
Z_m	0.08	0.10	0.21	0.53	0.12
H_d	0.09	0.06	0.13	0.55	0.27
C_b	0.08	0.05	0.09	0.51	0.20
L	0.08	1.13	2.26	0.50	0.01
B	0.08	0.45	1.06	0.57	0.03

Fig. 8: Results for the Dyke Cp case.

standarized regression coefficient and sobol indices, The variables more correlated and influential to the outputs C_p and S are: Q, H_d, Z_v, K_s and C_b.

Figures 8 and 9 present the results of this model for the variables Q, K_s, Z_v and Z_m. For the output C_p, the variables the three first variables present a clearer geometric pattern than the other inputs. In the case of the output S we recover as the most correlated variables as Q and Z_v. The other ones has not a clever pattern in which we could discriminate their influence.



Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	ρ^{Geom}	ρ^{SRC}
Q	0.08	4289.10	11554.51	0.63	0.65
K_s	0.09	80.24	163.61	0.51	0.41
Z_v	0.09	3.89	9.98	0.61	0.52
Z_m	0.09	4.77	9.46	0.50	0.08
H_d	0.11	3.01	5.13	0.41	0.32
C_b	0.09	2.35	4.71	0.50	0.22
L	0.09	46.66	83.16	0.44	0.01
B	0.09	23.11	45.34	0.49	0.03

Fig. 9: Results for the Dyke S case.

Also, the variable B has a higher value of ρ^{Geom} even if it is not correlated using the classic models. Figure 10 presents a zoom for this case. The reason comes from the bounding box which enclosed the manifold. The isolated points near to the box frontier creates 2-simplexes (triangles) far away of the concentrated data. Therefore, those simplexes rises artificially the area of the bounding box B. There are still improvements to be researched in order to create a robust version of the algorithm.

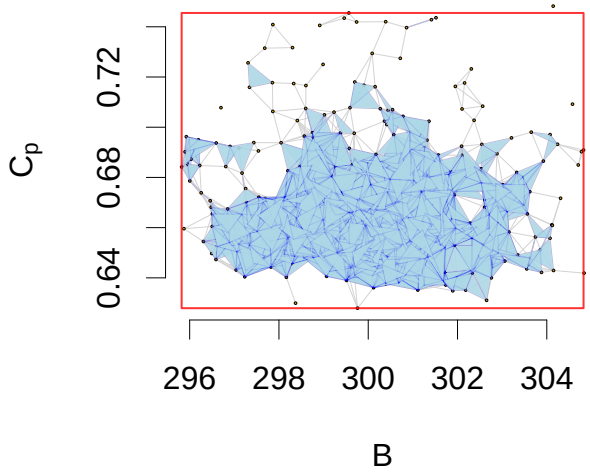


Fig. 10: Scaling issue with irrelevant variables. The VR-complex spreadout causes bounding boxes larger than the geometrical structure of the estimated manifold.

5 Conclusions and further research

As it was mentioned above, the aim of this paper was to build a sensitivity index that relied solely on the topological and geometrical features of a given data-cloud, since it is clear that purely analytic or statistic methods fail to recognize the structure within the projection of certain variables, primarily when the input is of zero sum, that is, what might be considered artificial noise. In such cases, those projections, or the variables in question, have positive conditional variance that might contribute to the model in ways that hadn't been explored so far.

Our index proved to be reliable in detecting this conditional variance when the variable is of zero sum, differentiating between pure random noise and well structured inputs -even of zero sum-. In the cases where the model presents pure noise our index coincides fully with other method's indexes in detecting relevant structured inputs, in the other cases our index reflects the presence of structure in all the variables, which was the case we wanted to explore.

As of sensitivity, we can not fully declare that our index measures it accurately, at least for the moment, and in a conventional way. To achieve the confidence for such a claim, we have identified a series of research problems to be dealt with in near future, namely: Improving the algorithm for the construction of the base graph, or change it completely to make the process more efficient and hence allow ourselves to run more sophisticated examples, both theoretical as well as real data examples from well known models. One of the central points to be discussed and

studied further is the determination of the radius of proximity, which we believe must be given by the data set itself, probably by a more detailed application of persistent homology. Finally we will be looking forward to extend our method to more than one variable at the time, to be able to check for crossed relevance.

We do not claim this to be an exhaustive list of problems related to the improvement of our method, but are confident that they would help us run more examples and more sophisticated ones, as well as will help us get more data to compare our results with other methods.

Acknowledgements

We acknowledge Santiago Gallón for enlightining discussions about the subject. His help has been very valuable.

First and second authors acknowledge the financial support from CIMPA, Centro de Investigaciones en Matemática Pura y Aplicada through the projects 821–B7–254 and 821–B8–221 respectively.

Third author acknowledges the financial support from CIMM, Centro de Investigaciones Matemáticas y Metamatemáticas through the project 820–B8–224.

The three authors also acknowledge Escuela de Matemática, Universidad de Costa Rica for their support.

6 Supplementary Material

R-package for TopSA routine: R-package "TopSA" estimates sensitivity indices reconstructing the embedding manifold of the data. The reconstruction is done via a Vietoris Rips with a fixed radius. Then the homology of order 2 and the indices are estimated.

References

Balasubramanian, M. (2002). The Isomap Algorithm and Topological Stability. *Science*, 295(5552):7a–7.

Barden, D. and Thomas, C. (2003). *An Introduction to Differential Manifolds*. Imperial College Press.

Bellman, R. (1957). *Dynamic Programming*. Dover Books on Computer Science Series. Princeton University Press.

Bellman, R. (1961). *Adaptive control processes: A guided tour*, volume 4 of *'Rand Corporation. Research studies*. Princeton University Press.

Bernstein, M., de Silva, V., Langford, J. C., and Tenenbaum, J. B. (2000). Graph approximations to geodesics on embedded manifolds. *Igarss 2014*, 01(1):1–5.

Bivand, R. and Rundel, C. (2018). *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with {R}, Second edition*. Springer, NY.

Borgonovo, E., Tarantola, S., Plischke, E., and Morris, M. D. (2014). Transformations and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):925–947.

Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Carlsson, G. (2014). Topological pattern recognition for point cloud data. *Acta Numerica*, 23(23):289–368.

Csárdi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*.

de Rocquigny, É. (2006). La maîtrise des incertitudes dans un contexte industriel. 1re partie : une approche méthodologique globale basée sur des exemples. *Journal de la société française de statistique*, 147(3):33–71.

Dimeglio, C., Gallón, S., Loubes, J. M., and Maza, E. (2014). A robust algorithm for template curve estimation based on manifold embedding. *Computational Statistics and Data Analysis*, 70:373–386.

Eppstein, D., Löffler, M., and Strash, D. (2010). Listing All Maximal Cliques in Sparse Graphs in Near-Optimal Time. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6506 LNCS, pages 403–414.

Gallón, S., Loubes, J.-M., and Maza, E. (2013). Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129–142.

Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75.

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, 32(2):135–154.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, volume 1 of *Springer Series in Statistics*. Springer-Verlag New York, New York, NY.

Hatcher, A. (2000). *Algebraic topology*. Cambridge Univ. Press, Cambridge.

Iooss, B. and Lemaître, P. (2015). A Review on Global Sensitivity Analysis Methods. In Dellino, G. and Meloni, C., editors, *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, pages 101–122. Springer US, Boston, MA.

Lang, S. (1963). *Introduction to differentiable manifolds*, volume 275.

Munkres, J. R. (1984). *Elements of Algebraic Topology*. Addison-Wesley.

Pebesma, E. and Bivand, R. (2005). Classes and methods for spatial data in R. *The Newsletter of the R Project*, 5(2):9–13.

R Core Team (2017). R: A Language and Environment for Statistical Computing.

Saltelli, A., Chan, K., and Scott, E. M. (2009). *Sensitivity Analysis*. Wiley, New York, 1 edition edition.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2007). *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd, Chichester, UK.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2002). *Sensitivity Analysis in Practice*. John Wiley & Sons, Ltd, Chichester, UK.

Snášel, V., Nowaková, J., Xhafa, F., and Barolli, L. (2017). Geometrical and topological approaches to Big Data. *Future Generation Computer Systems*, 67:286–296.

Sobol', I. M. (1993). Sensitivity Estimates for Nonlinear Mathematical Models. *Mathematical Modeling and Computational experiment*, 1(4):407–414.

Sobol', I. M. and Levitan, Y. (1999). On the use of variance reducing multipliers in Monte Carlo computations of a global sensitivity index. *Computer Physics Communications*, 117(1-2):52–61.

Tenenbaum, J. B. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323.

Wei, P., Lu, Z., and Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, 142:399–432.

Zomorodian, A. (2010). Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3):263–271.